Published online 6 December 2013 *Nucleic Acids Research, 2014, Vol. 42, Database issue* ***D749–D755***
doi:10.1093/nar/gkt1196

Ensembl 2014

Paul Flicek[1,2,*], M. Ridwan Amode[2], Daniel Barrell[2], Kathryn Beal[1], Konstantinos Billis[2], Simon Brent[2], Denise Carvalho-Silva[1], Peter Clapham[2], Guy Coates[2], Stephen Fitzgerald[1], Laurent Gil[1], Carlos García Girón[2], Leo Gordon[1], Thibaut Hourlier[2], Sarah Hunt[1], Nathan Johnson[1], Thomas Juettemann[1], Andreas K. Kähäri[2], Stephen Keenan[1], Eugene Kulesha[1], Fergal J. Martin[2], Thomas Maurel[1], William M. McLaren[1], Daniel N. Murphy[2], Rishi Nag[2], Bert Overduin[1], Miguel Pignatelli[1], Bethan Pritchard[2], Emily Pritchard[1], Harpreet S. Riat[2], Magali Ruffier[1], Daniel Sheppard[2], Kieron Taylor[1], Anja Thormann[1], Stephen J. Trevanion[2], Alessandro Vullo[1], Steven P. Wilder[1], Mark Wilson[2], Amonida Zadissa[1], Bronwen L. Aken[2], Ewan Birney[1], Fiona Cunningham[1], Jennifer Harrow[2], Javier Herrero[1], Tim J.P. Hubbard[2], Rhoda Kinsella[1], Matthieu Muffato[1], Anne Parker[2], Giulietta Spudich[1], Andy Yates[1], Daniel R. Zerbino[1] and Stephen M.J. Searle[2]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD and [2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

Received October 31, 2013; Accepted November 1, 2013

ABSTRACT

Ensembl (http://www.ensembl.org) creates tools and data resources to facilitate genomic analysis in chordate species with an emphasis on human, major vertebrate model organisms and farm animals. Over the past year we have increased the number of species that we support to 77 and expanded our genome browser with a new scrollable overview and improved variation and phenotype views. We also report updates to our core datasets and improvements to our gene homology relationships from the addition of new species. Our REST service has been extended with additional support for comparative genomics and ontology information. Finally, we provide updated information about our methods for data access and resources for user training.

INTRODUCTION

The Ensembl project (http://www.ensembl.org) creates and distributes genome annotations and provides integrated views of other valuable genomic data for supported chordate genomes. Our resources are intended to serve as community reference datasets on which other genomic research can be built. As such, Ensembl provides unique tools, datasets and user support compared to similar projects such as the UCSC Genome Browser (1), while supporting community standards that promote interoperability in genomics. For example, we have developed and distribute an extensive, open software infrastructure with diverse analysis pipelines supporting a variety of genome analyses (2) and the artificial intelligence inspired eHive analysis management system (3); data mining and analysis tools that include BioMart (4) and the Ensembl Variant Effect Predictor (VEP) (5); supported and robust application programing interfaces (APIs) (6) and a unique genome browser interface (7). Our software is distributed using a permissive Apache-style open-source license meaning that, unlike similar software, it is free for all potential users. Additionally, our data is provided without restriction and we have the most comprehensive suite of training options of any public genomics tool to maximize usability. In common with the UCSC Genome Browser, Ensembl supports community standard file formats such as BAM, BED, wiggle and other common file types. We have also incorporated support for track hubs over the past year to enable researchers to set up and view large-scale datasets. For example, the data produced by the ENCODE consortium (8) can be viewed by loading the ENCODE track hub (9) and users can then access an experiment matrix within the

*To whom correspondence should be addressed. Tel: +44 1223 492 581; Fax: +44 1223 494 494; Email: flicek@ebi.ac.uk

Ensembl configuration menu and quickly select datasets by cell or experiment type. However, over and above simply displaying these data, Ensembl uses them as described below in our integrative Regulatory Build analysis resulting in an evidence-based annotation of whole-genome regulation.

Ensembl resources are available for a total of 77 species as of release 73 (September 2013) with human, mouse, zebrafish, rat and various farm animals having the most extensive support. For 60 chordate species, we have full support comprising evidence-based gene annotation and comparative genomics analysis. In addition, for 18 of these species, there are variation resources and regulatory annotation for human and mouse. At present 13 additional chordate species are accessible with basic support via Ensembl preview sites (available from http://pre.ensembl.org), which provide BLAST access to the genome data and genome visualization, but not a complete gene build. Three non-chordate model species are also fully supported by Ensembl—worm (*Caenorhabditis elegans*), fruit fly (*Drosophila melanogaster*) and yeast (*Saccharomyces cerevisiae*)—with imported annotation from their respective genome databases in partnership with the Ensembl Genomes project (10). All fully supported species are accessible via the Ensembl BioMart, the Ensembl APIs and web displays. All data are also available for querying via our public MySQL servers, as full data downloads and as an Amazon public dataset.

Since our last report (11), we have added two new species with full gene annotation and comparative genomics support: duck (*Anas platyrhynchos*) (12) and collared flycatcher (*Ficedula albicollis*) and one new species with variation support: gibbon. New assemblies with corresponding updates to the gene annotations, alignments and variation data were also provided for rat, cat and chicken. At the same time, we added seven new species with basic support on the Ensembl preview site: blind cave fish (*Astyanax mexicanus*), white rhinoceros (*Ceratotherium simum simum*), baboon (*Papio anubis*), prairie vole (*Microtus ochrogaster*), vervet monkey (*Chlorocebus sabaeus*), naked mole-rat (*Heterocephalus glaber*) and aardvark (*Orycteropus afer*) and updated the preview sites for common shrew (*Sorex araneus*), bottle-nosed dolphin (*Tursiops truncatus*), American pika (*Ochotona princeps*) and armadillo (*Dasypus novemcinctus*) with new assemblies. In addition, as the human and mouse genome assemblies are updated regularly by the Genome Reference Consortium (GRC) to include alternate sequences in the form of 'fix' and 'novel' assembly patches (13), we include these additional alternate sequences and annotate them with genes, variation and other features as appropriate. Ensembl release 73 (September 2013) included the human GRCh37.p12 assembly (i.e. the twelfth patch release of the GRCh37 assembly) and the GRCm38.p1 mouse assembly.

In addition to the newly support species and community standards reported above, the most important updates over the last year that have advanced the project since our last report (11) include new and more comprehensive phenotype annotations most valuable to those interested in human disease research, scrollable genome browsing designed to appeal to all users of our web interface and new REST endpoints supporting more flexible analysis options for those users that interact with the Ensembl resources programmatically. These and other features are described in more detail below.

Ensembl browser

This year we significantly updated Ensembl's main Region in Detail page with the full incorporation of the Javascript-based, scrollable and zoomable browser, Genoverse, in place of the overview panel that had been a part of Ensembl for >10 years. Older, unsupported browsers fall back to the previous non-scrolling overview image. Genoverse allows users to scroll back and forth along the genome and update the main image below it to show the new region (Figure 1A). Our search engine was also upgraded from Lucene to Solr and we implemented a new search interface with features such as faceting and auto-completion.

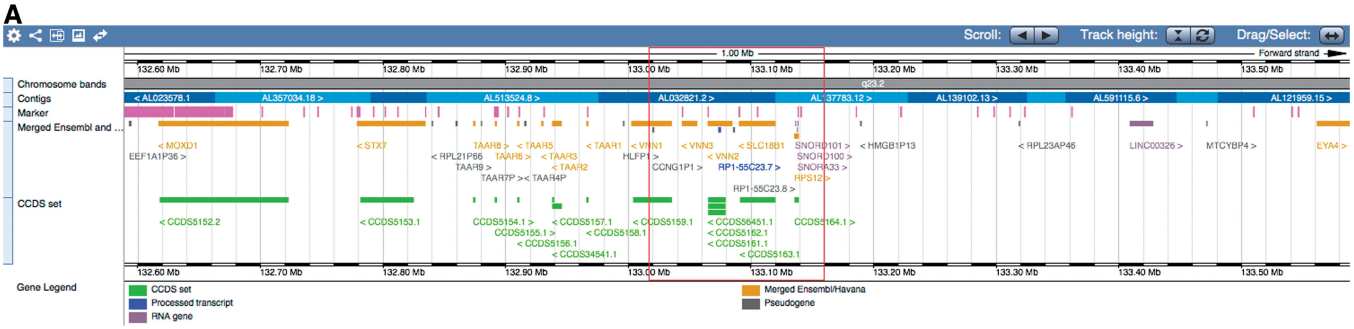
Our web displays dedicated to variation and phenotype data were also markedly improved. We specifically focused on displays for structural variants, which are now coloured by class and the higher quality structural variants from the 1000 Genomes Project are provided in a separate track. We introduced a page for structural variant phenotype data and now have additional phenotype data, including the variants from NCBI's ClinVar project that are classified as being probable-pathogenic, pathogenic, drug-response or histocompatibility. Phenotype data from multiple sources are integrated and displayed for relevant genes (Figure 1B). Variants in regulatory regions are now annotated on all tracks and variant names are visible when zoomed in on all displays. We have also improved the VEP visual output in the form of summary pie charts.

Beyond these major developments, there have been other important improvements. In particular, we improved the handling of user data with a streamlined upload interface and support for uploading VEP output files. Additionally, configuration of complex data hubs has been made easier by displaying the track options in a matrix similar to the existing configurations for regulation data.

Ensembl annotations

All Ensembl annotations whether gene, variation or regulation, are based on integration of relevant data sources. We update the human gene set for every Ensembl release via a merge of the Ensembl evidence-based automatic annotation and Havana (14) manual annotation to produce an updated GENCODE gene set. This set also includes all current human Consensus Coding Sequence (CCDS) gene models (15). Manual annotation from Havana is additionally incorporated into our gene sets on alternate releases for zebrafish (16) and for mouse, which also includes all current CCDS gene models. Pig includes manual annotation from Havana on selected regions of the genome. The year 2013 has seen the inclusion of RNASeq data for seven species: human, chicken, cat, collared flycatcher,

A



B

Gene: ATP2A2 ENSG00000174437

Description	ATPase, Ca++ transporting, cardiac muscle, slow twitch 2 [Source:HGNC Symbol;Acc:812]
Location	Chromosome 12: 110,718,561-110,788,898 forward strand.
INSDC coordinates	chromosome:GRCh37:CM000674.1:110718561:110788898:1
Transcripts	This gene has 13 transcripts (splice variants) Show transcript table

Phenotype

List of phenotype(s) associated with the gene ENSG00000174437

Filter

Source	Phenotype	Locations
DDG2P	acrokeratosis verruciformis (AKV)	View on Karyotype
DDG2P	Darier disease (DD)	View on Karyotype
OMIMGENE	ACROKERATOSIS VERRUCIFORMIS; AKV	View on Karyotype
OMIMGENE	DARIER-WHITE DISEASE; DAR	View on Karyotype
Orphanet	Acrokeratosis verruciformis of Hopf	View on Karyotype
Orphanet	DARIER DISEASE	View on Karyotype

Phenotypes associated with the gene from variation annotations

Show 10 entries Filter

Number of variants	Show/hide details	Phenotype	Locations	Biomart	Source(s)
1	Show	ACROKERATOSIS VERRUCIFORMIS	View on Karyotype	-	OMIM
214	Show	ALL variations with a phenotype annotation			-
148	Show	Annotated by HGMD but no phenotype description is publicly available	-	-	HGMD-PUBLIC
1	Show	COSMIC:tumour_site:autonomic_ganglia	View on Karyotype	View list in BioMart	COSMIC
3	Show	COSMIC:tumour_site:breast	View on Karyotype	View list in BioMart	COSMIC
5	Show	COSMIC:tumour_site:endometrium	View on Karyotype	View list in BioMart	COSMIC
4	Show	COSMIC:tumour_site:haematopoietic_and_lymphoid_tissue	View on Karyotype	View list in BioMart	COSMIC
4	Show	COSMIC:tumour_site:kidney	View on Karyotype	View list in BioMart	COSMIC
21	Show	COSMIC:tumour_site:large_intestine	View on Karyotype	View list in BioMart	COSMIC
8	Show	COSMIC:tumour_site:lung	View on Karyotype	View list in BioMart	COSMIC

Showing 1 to 10 of 21 entries << < 1 2 3 > >>

Figure 1. (**A**) The scrollable Genoverse view (with view control icons in the upper right corner) provides the overview panel on the Region in Detail page. Image from URL: http://e73.ensembl.org/Homo_sapiens/Location/View?r=6:133017695-133161157. (**B**) Phenotype data from DDG2P, OMIM, Orphanet, HGMD and COSMIC for the human gene ATP2A2. Image created from URL http://e73.ensembl.org/Homo_sapiens/Gene/Phenotype?db=core;g=ENSG00000174437;r=12:110718561-110788898.

gibbon, rabbit and anole lizard. For gibbon, rabbit and anole lizard, the RNASeq data were used to update an existing standard gene annotation whereas for other species (except human) the data were integrated into the annotation as part of the primary gene-build process. Some of these species are provided with tissue-specific RNASeq samples which allow users to explore tissue-specific expression.

Ensembl variation annotation integrates all publicly available variation datasets to provide a coherent and complete resource for variome interpretation across 217 million variants in the 18 Ensembl species with supported variation resources. Basic variation data including genomic location, allele changes, allele and genotype frequencies and population data are imported for SNPs and indels from dbSNP (17) and for structural variants

from DGVa (18). Additional human data are imported directly from the 1000 Genomes Project (19), the Exome Sequencing Project (20) and from 14 individual genomes that provide genotype information. In addition to the newly supported gibbon data listed above, over the past year, updated variation resources were released for human, platypus, cow, mouse, pig, zebrafish, opossum, orang-utan and macaque. We also extended cross-references to new and popular genotyping chips for human, chicken, horse and cow.

We significantly expanded our support for human phenotype data in Ensembl beyond the UniProt (21), OMIM (22), EGA, HGMD Public (23) (variation location only), COSMIC somatic mutations (24) and NHGRI GWAS Catalog (25) resources that we have supported in the past. New data from ClinVar, Orphanet (26), the Developmental Disorder Genotype—Phenotype Database (DDG2P) from DECIPHER (27), dbGaP, Phencode and the MAGIC and GIANT consortiums have now been fully integrated into Ensembl (Figure 1B). From these data sources, we select only the significant associations to display on the website and provide full datasets in the database and BioMart. For variants stored in LOVD (28), Ensembl queries LOVD directly and displays the information on the appropriate variation web page. In addition, we have now incorporated phenotype information for other species. Mouse phenotype data are provided from international projects including EuroPhenome (29) and IMPC (International Mouse Phenotyping Consortium) (30). For other animals, data are imported from the Online Mendelian Inheritance in Animals (OMIA) database (31). This year we have also developed a new pipeline to cross-reference publications citing variants from EuropePMC, NCBI and UCSC.

Regulatory annotation in Ensembl is currently available across multiple human and mouse cell lines. The main resource, the Ensembl Regulatory Build is a comprehensive synthesis of functional assays provide by a number of consortia, such as ENCODE (8) and Roadmap Epigenomics Mapping Consortium (32). Although the raw data from these projects can be displayed directly on Ensembl through dedicated track hubs, the Regulatory Build is a higher level integrated analysis that defines a collection of Regulatory Features (i.e. regions of the genome that display regulatory activity in one of 13 human or five murine cell lines). Where relevant, transcription-factor binding sites are predicted on these regions using the JASPAR binding motifs (33).

Additionally, Ensembl links out to relevant externally curated databases of regulatory data including enhancer regions from VISTA (34), miRNA binding sites from Microcosm http://www.ebi.ac.uk/enright-srv/microcosm/) and eQTLs from Genevar (35). Several reference DNA methylation experiments are also included.

COMPARATIVE GENOMICS RESOURCES

Whole-genome alignments of vertebrate species are provided within the Ensembl Compara database. Because all genome assemblies are not sequenced to the same level of completeness, we group the assemblies into two tiers for differential processing. High quality genomes from 13 species are aligned into a progressive multiple sequence alignment using the EPO (Enredo-Pecan-Ortheus) pipeline (36,37), which also estimates the underlying ancestral genome sequences. The low coverage genomes of an additional 23 species, that are much more fragmented, are inserted into the previous alignment by mapping them onto the human assembly with LASTZ. We also produce clade-specific multiple sequence alignments for primates, birds, fish and amniotes. In particular, the fish multiple alignment has been extended to eight species this year. From these alignments, we compute the conservation at every position, using GERP (38).

Our gene-based comparative genomics resources are updated every release to incorporate new species, updated assemblies and gene annotation sets. These include gene phylogenetic trees, gene families and gene dynamics. This year, the inclusion of duck and collared flycatcher and an update of the guide species tree have greatly improved the quality of the gene trees in the Sauria clade and reduced the number of poorly supported duplications from 74% to 30%. In close collaboration with the TreeFam (39) and Ensembl Genomes (10) projects, we will migrate to an HMM-based classification for GreeTree annotation, which will reduce a key quadratic complexity to a linear one.

Data access, data mining and quality control

Ensembl's REST service, available at http://beta.rest.ensembl.org, continues to be actively developed as a public beta (Figure 2). This year has seen the addition of 12 new endpoints including access to translation features, SNVs and protein domains, as well as access to whole-genome alignments by region. Additionally, the REST API is now able to query GeneTrees by their containing member stable identifier or gene symbols such as HGNC. Other new endpoints include the ability to use a stable identifier to identify overlapping features and location information as well as access to NCBI taxonomy and ontology datasets. The ontology endpoints currently provide the gene ontology (GO) (40), sequence ontology (SO) (41) and experimental factor ontology (EFO) (42) information used within Ensembl. The REST service will move out of beta during the next year coinciding with the introduction of POST requests and improved VEP integration.

The more established BioMart data-mining tool (43) provides users with a variety of ways of accessing the Ensembl data quickly and with relative ease. Users can choose to access the data via the MartView web interface or via the MartService routes including the BioMart Perl API, DAS server, SOAP, REST or BioConductor biomaRt package. The Ensembl BioMart databases (4) are built from scratch each release in order to incorporate the latest annotated and imported data, and they are current with the data resources described above.

Beyond programmatic and tool-based data-access methods, we continue to provide complete data downloads

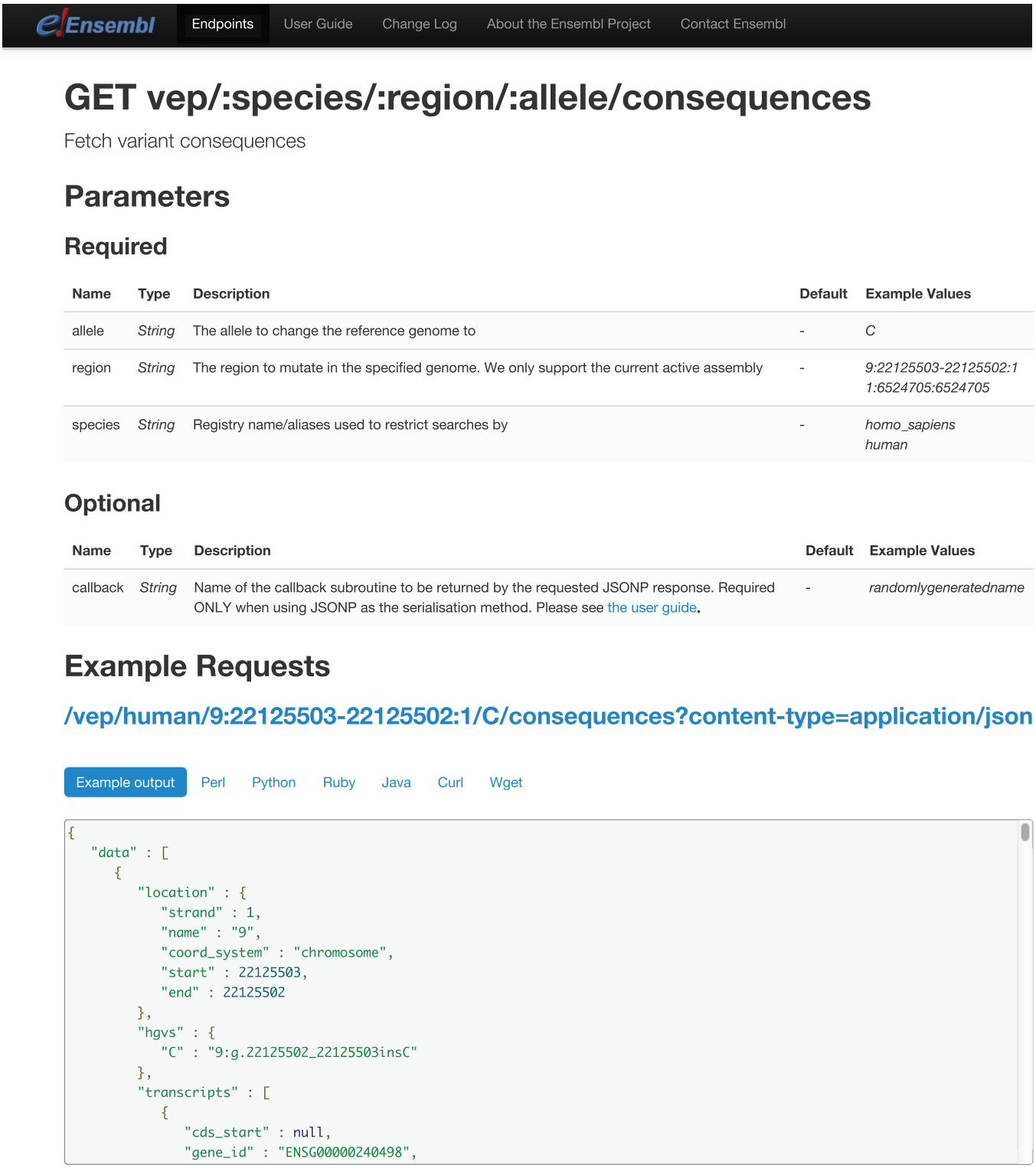


Figure 2. Usage and example output for the Ensembl REST server Fetch Variant Consequences endpoint.

in a variety of formats including the VCF files that were introduced this year to distribute many subsets of the Ensembl variation data.

To manage the increasing size and complexity of Ensembl releases, we have increased our quality control (QC) procedures over the past year. These are an essential part of each release cycle and range from validation testing of the various APIs to methods for checking data and database integrity. The Ensembl gene set is also independently analyzed using a specific curation/QC pipeline

run for all updates of the human and mouse gene sets. This procedure compares the set of Ensembl translations for a particular species directly to the publicly available sequence resources UniProt (21) and RefSeq (44) and reports the percentage identity of the alignments. In addition to the above species, the pipeline has been employed for the rat, zebrafish and chicken genomes.

Ensembl variation and regulatory resources also rely on comprehensive and flexible infrastructure to manage the growing amount of relevant datasets in the public domain. In preparation for the updated human reference genome (GRCh38) expected at the end of 2013, we consolidated these pipelines to work automatically over a large number of files with minimal supervision. Specifically, more of our pipelines are run using the eHive system (3) and employ both a modular structure to the analysis and on-the-fly calculation for specific data types stored in the databases.

Outreach and training

Ensembl supports users through worldwide face-to-face training workshops, our helpdesk@ensembl.org and dev@ensembl.org email lists, online training and social media. This year we made several changes to intensify and connect our user interactions on Twitter (https://twitter.com/Ensembl), Facebook (https://www.facebook.com/Ensembl.org) and the Ensembl Blog (http://www.ensembl.info/). Together these methods have proven extremely effective for supporting users beyond the traditional mailing list and FAQ model that we continue to maintain.

Distance and on-line training are provided through our Helpdesk channel on YouTube, which saw >70% growth in the number of subscribers and incorporated two new videos during the year. For those users wanting more intensive training, we filmed and now provide a one-day Ensembl browser workshop (http://www.ebi.ac.uk/training/online/course/ensembl-filmed-browser-workshop) and a three-day Ensembl API workshop (http://www.ebi.ac.uk/training/online/course/ensembl-filmed-api-workshop), both complete with videos and exercises. A Quick Ensembl course providing a short introduction to the browser (http://www.ebi.ac.uk/training/online/course/ensembl-quick-tour-0) complements the more complete workshops.

ACKNOWLEDGEMENTS

The authors wish to thank all of their users, the systems teams who maintain their computational infrastructure and those researchers who have provided data to Ensembl in advance of publication under the understandings of the Fort Lauderdale meeting discussing Community Resource Projects and the Toronto meeting on pre-publication data sharing.

FUNDING

The Wellcome Trust provides majority funding for the Ensembl project [WT095908 and WT098051] with additional funding for specific project components from the National Human Genome Research Institute [U41HG007234, 1R24RR032658 and 1R01HD074078], the BBSRC [BB/I025506/1, BB/I025360/1 and BB/K009524/1], the European Molecular Biology Laboratory and as specified: The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 222664. (''Quantomics''); This Publication reflects only the author's views and the European Community is not liable for any use that may be made of the information contained herein. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number [200754] - the GEN2PHEN project; The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement [n° 282510 - BLUEPRINT]; Work supported by the European Commission within the Framework Programme 7 Capacities Specific Programme, under Grant Agreement [no. 312301] (Helix Nebula - The Science Cloud; Rat genomics resources receive additional support as specified: The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement [N° HEALTH-F4-2010-241504] (EURATRANS). Funding for open access charge: The Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Meyer,L.R., Zweig,A.S., Hinrichs,A.S., Karolchik,D., Kuhn,R.M., Wong,M., Sloan,C.A., Rosenbloom,K.R., Roe,G., Rhead,B. *et al.* (2013) The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res.*, **41**, D64–D69.
2. Potter,S.C., Clarke,L., Curwen,V., Keenan,S., Mongin,E., Searle,S.M.J., Stabenau,A., Storey,R. and Clamp,M. (2004) The Ensembl analysis pipeline. *Genome Res.*, **14**, 934–941.
3. Severin,J., Beal,K., Vilella,A.J., Fitzgerald,S., Schuster,M., Gordon,L., Ureta-Vidal,A., Flicek,P. and Herrero,J. (2010) eHive: an artificial intelligence workflow system for genomic analysis. *BMC Bioinform.*, **11**, 240.
4. Kinsella,R.J., Kähäri,A., Haider,S., Zamora,J., Proctor,G., Spudich,G., Almeida-King,J., Staines,D., Derwent,P., Kerhornou,A. *et al.* (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford)*, **2011**, bar030.
5. McLaren,W., Pritchard,B., Rios,D., Chen,Y., Flicek,P. and Cunningham,F. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, **26**, 2069–2070.
6. Stabenau,A., McVicker,G., Melsopp,C., Proctor,G., Clamp,M. and Birney,E. (2004) The Ensembl core software libraries. *Genome Res.*, **14**, 929–933.
7. Parker,A., Bragin,E., Brent,S., Pritchard,B., Smith,J.A. and Trevanion,S. (2010) Using caching and optimization techniques to improve performance of the Ensembl website. *BMC Bioinform.*, **11**, 239.
8. ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
9. Rosenbloom,K.R., Sloan,C.A., Malladi,V.S., Dreszer,T.R., Learned,K., Kirkup,V.M., Wong,M.C., Maddren,M., Fang,R., Heitner,S.G. *et al.* (2013) ENCODE Data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res.*, **41**, D56–D63.

10. Kersey,P.J., Staines,D.M., Lawson,D., Kulesha,E., Derwent,P., Humphrey,J.C., Hughes,D.S.T., Keenan,S., Kerhornou,A., Koscielny,G. *et al.* (2012) Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Res.*, **40**, D91–D97.
11. Flicek,P., Ahmed,I., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**, D48–D55.
12. Huang,Y., Li,Y., Burt,D.W., Chen,H., Zhang,Y., Qian,W., Kim,H., Gan,S., Zhao,Y., Li,J. *et al.* (2013) The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. *Nat Genet.*, **45**, 776–783.
13. Church,D.M., Schneider,V.A., Graves,T., Auger,K., Cunningham,F., Bouk,N., Chen,H.-C., Agarwala,R., McLaren,W.M., Ritchie,G.R.S. *et al.* (2011) Modernizing reference genome assemblies. *PLoS Biol.*, **9**, e1001091.
14. Wilming,L.G., Gilbert,J.G.R., Howe,K., Trevanion,S., Hubbard,T. and Harrow,J.L. (2008) The vertebrate genome annotation (Vega) database. *Nucleic Acids Res.*, **36**, D753–D760.
15. Harte,R.A., Farrell,C.M., Loveland,J.E., Suner,M.-M., Wilming,L., Aken,B., Barrell,D., Frankish,A., Wallin,C., Searle,S. *et al.* (2012) Tracking and coordinating an international curation effort for the CCDS Project. *Database (Oxford)*, **2012**, bas008.
16. Howe,K., Clark,M.D., Torroja,C.F., Torrance,J., Berthelot,C., Muffato,M., Collins,J.E., Humphray,S., McLaren,K., Matthews,L. *et al.* (2013) The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, **496**, 498–503.
17. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
18. Lappalainen,I., Lopez,J., Skipper,L., Hefferon,T., Spalding,J.D., Garner,J., Chen,C., Maguire,M., Corbett,M., Zhou,G. *et al.* (2013) dbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res.*, **41**, D936–D941.
19. 1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
20. Fu,W., O'Connor,T.D., Jun,G., Kang,H.M., Abecasis,G., Leal,S.M., Gabriel,S., Altshuler,D., Shendure,J., Nickerson,D.A. *et al.* (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, **493**, 216–220.
21. UniProt Consortium. (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.*, **41**, D43–D47.
22. Amberger,J., Bocchini,C. and Hamosh,A. (2011) A new face and new challenges for Online Mendelian Inheritance in Man (OMIM(®)). *Hum Mutat.*, **32**, 564–567.
23. Stenson,P.D., Ball,E.V., Mort,M., Phillips,A.D., Shiel,J.A., Thomas,N.S.T., Abeysinghe,S., Krawczak,M. and Cooper,D.N. (2003) Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.*, **21**, 577–581.
24. Forbes,S.A., Bindal,N., Bamford,S., Cole,C., Kok,C.Y., Beare,D., Jia,M., Shepherd,R., Leung,K., Menzies,A. *et al.* (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.*, **39**, D945–D950.
25. Manolio,T.A., Collins,F.S., Cox,N.J., Goldstein,D.B., Hindorff,L.A., Hunter,D.J., McCarthy,M.I., Ramos,E.M., Cardon,L.R., Chakravarti,A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
26. Rath,A., Olry,A., Dhombres,F., Brandt,M.M., Urbero,B. and Ayme,S. (2012) Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. *Hum. Mutat.*, **33**, 803–808.
27. Swaminathan,G.J., Bragin,E., Chatzimichali,E.A., Corpas,M., Bevan,A.P., Wright,C.F., Carter,N.P., Hurles,M.E. and Firth,H.V. (2012) DECIPHER: web-based, community resource for clinical interpretation of rare variants in developmental disorders. *Hum Mol Genet*, **21**, R37–R44.
28. Fokkema,I.F.A.C., Taschner,P.E.M., Schaafsma,G.C.P., Celli,J., Laros,J.F.J. and den Dunnen,J.T. (2011) LOVD v.2.0: the next generation in gene variant databases. *Hum. Mutat.*, **32**, 557–563.
29. Morgan,H., Beck,T., Blake,A., Gates,H., Adams,N., Debouzy,G., Leblanc,S., Lengger,C., Maier,H., Melvin,D. *et al.* (2010) EuroPhenome: a repository for high-throughput mouse phenotyping data. *Nucleic Acids Res.*, **38**, D577–D585.
30. Mallon,A.-M., Iyer,V., Melvin,D., Morgan,H., Parkinson,H., Brown,S.D.M., Flicek,P. and Skarnes,W.C. (2012) Accessing data from the International Mouse Phenotyping Consortium: state of the art and future plans. *Mamm. Genome*, **23**, 641–652.
31. Lenffer,J., Nicholas,F.W., Castle,K., Rao,A., Gregory,S., Poidinger,M., Mailman,M.D. and Ranganathan,S. (2006) OMIA (Online Mendelian Inheritance in Animals): an enhanced platform and integration into the Entrez search interface at NCBI. *Nucleic Acids Res.*, **34**, D599–D601.
32. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F., Ren,B., Milosavljevic,A., Meissner,A., Kellis,M., Marra,M.A., Beaudet,A.L., Ecker,J.R. *et al.* (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.*, **28**, 1045–1048.
33. Portales-Casamar,E., Thongjuea,S., Kwon,A.T., Arenillas,D., Zhao,X., Valen,E., Yusuf,D., Lenhard,B., Wasserman,W.W. and Sandelin,A. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105–D110.
34. Visel,A., Minovitsky,S., Dubchak,I. and Pennacchio,L.A. (2007) VISTA Enhancer Browser–a database of tissue-specific human enhancers. *Nucleic Acids Res.*, **35**, D88–D92.
35. Yang,T.-P., Beazley,C., Montgomery,S.B., Dimas,A.S., Gutierrez-Arcelus,M., Stranger,B.E., Deloukas,P. and Dermitzakis,E.T. (2010) Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics*, **26**, 2474–2476.
36. Paten,B., Herrero,J., Beal,K., Fitzgerald,S. and Birney,E. (2008) Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.*, **18**, 1814–1828.
37. Paten,B., Herrero,J., Fitzgerald,S., Beal,K., Flicek,P., Holmes,I. and Birney,E. (2008) Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.*, **18**, 1829–1843.
38. Cooper,G.M., Stone,E.A., Asimenos,G., NISC Comparative Sequencing Program, Green,E.D., Batzoglou,S. and Sidow,A. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15**, 901–913.
39. Ruan,J., Li,H., Chen,Z., Coghlan,A., Coin,L.J.M., Guo,Y., Hériché,J.-K., Hu,Y., Kristiansen,K., Li,R. *et al.* (2008) TreeFam: 2008 Update. *Nucleic Acids Res.*, **36**, D735–D740.
40. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.*, **25**, 25–29.
41. Eilbeck,K., Lewis,S.E., Mungall,C.J., Yandell,M., Stein,L., Durbin,R. and Ashburner,M. (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.*, **6**, R44.
42. Malone,J., Holloway,E., Adamusiak,T., Kapushesky,M., Zheng,J., Kolesnikov,N., Zhukova,A., Brazma,A. and Parkinson,H. (2010) Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, **26**, 1112–1118.
43. Smedley,D., Haider,S., Ballester,B., Holland,R., London,D., Thorisson,G. and Kasprzyk,A. (2009) BioMart–biological queries made easy. *BMC Genom.*, **10**, 22.
44. Pruitt,K.D., Tatusova,T., Brown,G.R. and Maglott,D.R. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.